Impact of CMOS Technology Scaling on Leakage Reduction Techniques

Saad Arrabi, Taeyoung Kim
ECE 6332 – Fall 2009
University of Virginia
{arrabi, tk2np}@virginia.edu

ABSTRACT

CMOS leakage power consumption has become a big problem recently. While technology is scaling down, the leakage power consumption is growing exponentially. In order to solve this problem, many researchers suggested some leakage reduction techniques. We consider some of those techniques, and test them to see if they are adequate for the current and future technologies. We looked at sleep and stack techniques. We apply these techniques to four kinds of CMOS technologies (65nm~22nm). We analyzed their performance for generic type of circuits and under multiple configurations.

1. INTRODUCTION

These days, researchers have been trying to figure out a way to design an ultra low-power circuit. Many portable and wireless devices that require low power are being made. These low-power circuit designs are based on a low-power CMOS chips. To produce such circuits, many low power techniques have to be used.

In the past, only dynamic power was the main issue of energy consumption. Technology though is in continuous evolving, and new problems began to be discovered. One of those issues is static leakage power. This problem is contributing to the overall power consumption. With technology shrinking, supply voltage and threshold voltage go down as well. The sub-threshold leakage and gate-oxide leakage power, however, are increasing exponentially.

In this paper, we are investigating some of the leakage reduction techniques proposed in the past and have been used in the present. We apply those techniques on cutting edge and future technologies to see the effectiveness for future circuit designs. We analyzed the characteristic of those techniques and evaluated their performances on several measured metrics.

This paper is organized as follows. Section 2 will provide background information on current leakage reduction techniques. Section 3 explains our method. In section 4 we present our results and our analysis. Section 5 concludes the paper.

2. BACKGROUND

As we mentioned earlier, we used couple of techniques to the state of art technology and some predictive technology models. We are exploring two main techniques and some of their variations sleep and stack techniques.

The basic principle of sleep technique is to turn off the circuit when the circuit is in standby mode [6]. To configure the circuit additional transistors are inserted at the header, footer, or both. The signal controlling them is separate sleep signal. Significant amounts of energy are able to be saved in standby mode because sleep transistor disconnects from the supply voltage or ground from the circuit. However, this technique could have some drawbacks because there is some overhead area as well some impact on dynamic operation of the circuit. The state of the cell tends to be lost due to broken pull-up and pull-down network as well.

Another similar technique using sleep transistors is the zigzag technique [2]. This technique is proposed to reduce the drawback of sleep technique, which is having the overhead delay for waking up from sleep mode. Although this technique is able to maintain the state of the circuit to prevent a floating state, a complex set of configurations are needed.

Another leakage reduction technique is cloning (stacking) each transistor from the pull up and pull down network [5]. Using stack technique, the state will be maintained. The base idea of the stack is that at least two or more transistors, stacked at each turned off network which reduces the leakage power. The stack technique may increase the internal resistance and the capacitance which can impact the dynamic operation significantly.

Sleepy-stack technique is on of the latest techniques for ultra low-power design [1]. The combination of sleep transistor and stack transistor is used for this approach. This technique is used only when ultra low-power mode is required and area is not constrained. This technique can maintain the state of the circuit better than stack or sleep techniques. This technique, as well, has large impact on dynamic operations. In this paper we will focus on sleep and stack techniques.

3. APPROACH

We chose the following techniques to analyze, detailed description of each technique is provided in 3.1.1 and 3.1.2.

- Shared footer sleep transistor
- Non-shared footer sleep transistor
- Stack with headers and footer
- Stack with only headers
- Stack with only footers

Each technique was simulated with five different widths starting from minimum width to five times the minimum width.

3.1 Leakage Techniques

3.1.1 Sleep Technique

In this technique, a NMOS transistor is added to the footer of the circuit or a PMOS transistor to the header of the circuit. This transistor is controlled by a separate signal. When the circuit is not operational (sleep mode) then the sleep transistor will be shut off. This will cut the direct link of the circuit to V_{SS} or V_{DD} which in return reduces the leakage current.

The sleep transistor can vary in width depending on how much delay the circuit can tolerate. Increase of the width will result in shorter dynamic delay, but will also increase the leakage in the circuit. Dynamic energy is not much affected by sleep transistor

since it is not driven by the same signals like in the stack technique.

Another variation we tried in our simulations is having one sleep transistor for the whole circuit or add one at the footer of each gate. Sharing same sleep transistor can be helpful in case some parts of the circuit will be non-active while the rest of the circuit is active.

The main benefits of this technique are its low impact on the dynamic delay and on the dynamic energy since the sleep transistor will be a simple passing transistor in active mode. The main downside of this technique is its inability to reduce leakage while the circuit is in active mode. This can prevent the amount of leakage reduction. In addition, the sleep transistor technique requires a separate sleep signal. This can add significant amount of complexity, especially if the circuit is active in unpredictable behavior.

3.1.2 Stack Technique

For this technique we add two transistors in the CMOS gates for each input. This will simply stack at least two identical transistors at the pull up and the pull down network in series. All the additional transistors are controlled by the same inputs of the gate. This way any leakage current have to go through two turned off transistor at least. This fact reduces the leakage current drastically. For our analysis we tried having the transistor at both, the headers and the footers and we tried having them at only one location at a time.

The main benefit of this technique is its ability of reducing leakage even in active mode, unlike the sleep technique which requires the circuit to go to sleep. This, however, comes at a cost of increase of the dynamic energy and delay. The cost of this technique can be significant since it is effectively doubling the gates each signal is required to drive.

3.2 Simulated Circuit

Since the leakage reduction techniques we chose are very generic and can be used in most types of circuits, we did our performance analysis using a random logical circuit we constructed to mimic the behavior of typical static logic circuits. The random circuit consists mainly of three paths of mainly inverters which link to each other NOR and NAND gates throughout the path.

The circuit we chose represents some characteristics of typical circuits like glitching. While this circuit does not identically emulate other circuits, it does provide similar trend which can be used to other circuits.

3.3 Measured Metrics

While many metrics and cost functions can be considered to evaluate the performance of a specific leakage reduction technique, we chose the following set of metrics to be our comparison categories. The metrics we chose cover most of the important aspects typical designers look for.

3.3.1 Dynamic Energy

For this metric, we measured the average current through the voltage supply rails over 16ns and 4 switches of the input, two 0 to 1 transition and two 1 to 0 transition. We subtracted the average leakage current from the result then we normalized the data to the base case.

3.3.2 Dynamic Delay

The metric was measured by finding the time the input reached the 50% of V_{DD} point of the transition to the 50% of V_{DD} point of the output signal. Then we averaged the transition between 0 to 1 and 1 to 0 to provide a more generic result.

3.3.3 Leakage Power

This metric was measured by measuring the leakage through the voltage supply during a static input. Then the results were averaged between the input 1 and the input 0. The sleep transistors were assumed to be in sleep mode (deactivated). For leakage through the sleep transistors in the active mode we assumed it will be equal to the base case.

4. EXPERIMENTAL RESULTS

The leakage in the future is predicted to increase exponentially (see Figure 1). The leakage is expected to jump drastically once the technology becomes small enough so the current will start leaking through the gates which will produce huge amount of additional leakage current. At 22 nm technologies, the predicted technology model predicts the leakage through circuits to be around 50 times the amount of leakage in the 65 nm technology.

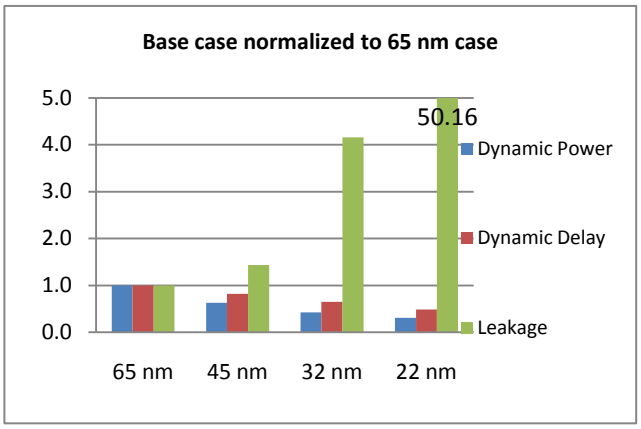

Figure 1: Leakage estimate for future technologies

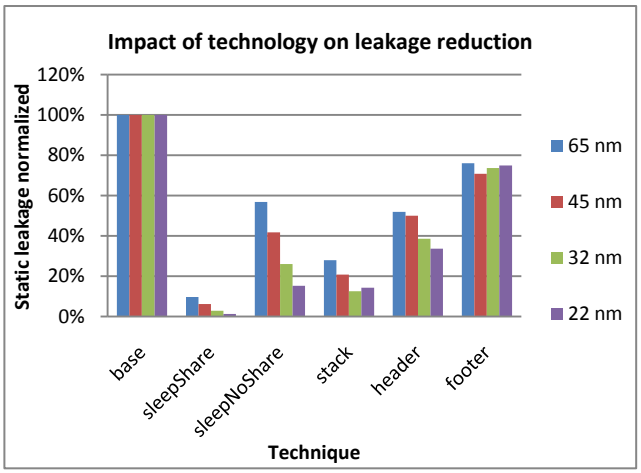

Figure 2: Leakage reduction techniques performance on future technologies

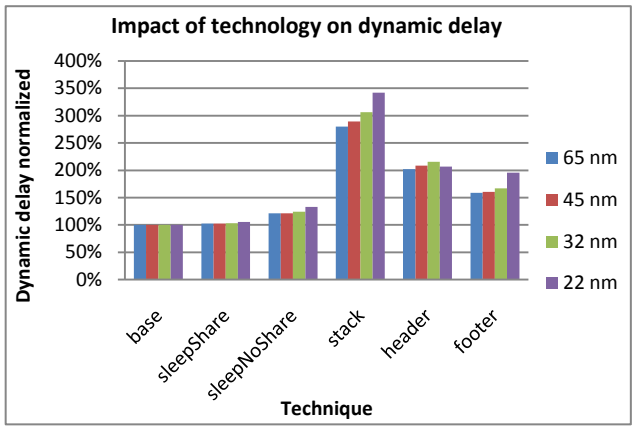

Figure 3: Dynamic delay impact of leakage reduction technique in the future

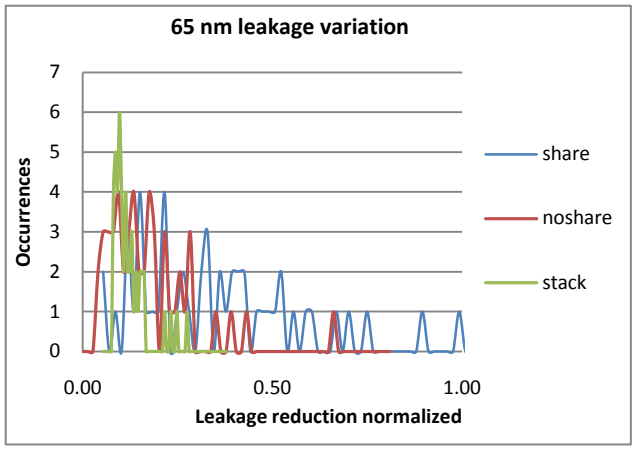

Figure 4: Leakage variation

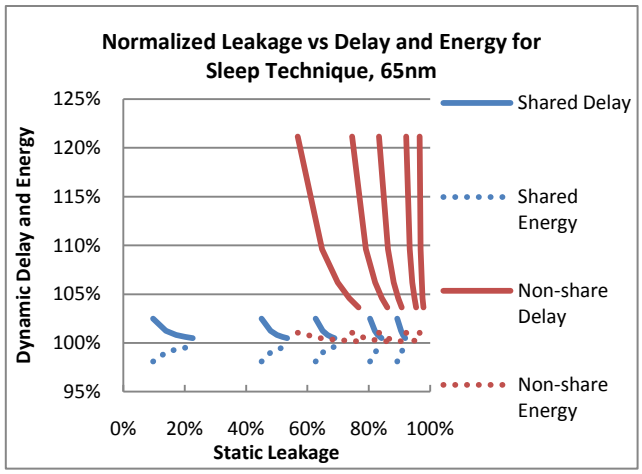

Figure 5: Sleep technique performance and impact

In Figure 2, we can see the how the leakage reduction technique is predicted to become more efficient in reducing the leakage. While the leakage reduction is anticipated to be more efficient, the impact of the leakage reduction techniques is getting slightly worse (see Figure 3). This suggests that leakage reduction techniques can be adjusted to maintain the same leakage percentage and have the same impact of the techniques on dynamic delay and energy we are seeing in current circuits.

Since the leakage energy will increase very faster and the dynamic energy is decreasing with technology scaling, the impact of the leakage reduction techniques will become less significant. This will enable more drastic techniques which can affect the dynamic energy drastically to become viable in the future technologies.

We see in the Figure 4 the coping ability of the leakage reduction techniques to the manufacturing variations. The variation will become more significant with the technology scaling. We can see how the stack and the non-shared sleep techniques are significantly more resilient to variation than the shared sleep technique. This is mainly due to the fact that shared sleep technique only uses one transistor which can vary in Gaussian behavior unlike the stack and the non-share sleep techniques which use many more transistors.

Since sleep technique can only reduce leakage current in sleep mode, we plotted the amount to be saved if the circuit is in sleep mode 100%, 60%, 40%, 20% and 10% of the time. We can see how the sleep technique has relatively small impact on dynamic delay and energy (see Figure 5). This occur mainly due to the fact that the sleep transistor is controlled by a separate signal, unlike the stack technique where the same input signal will have to drive additional gates.

We see, however, how the leakage current is increased significantly once we reduce the percentage the circuit is in sleep mode. This will limit the efficiency of this technique to only parts of the system that is not active most of the time.

In addition, the effect of changing the width of the sleep transistor is shown. The wider the sleep transistor is the less dynamic delay and energy impact on the circuit but of course it increases the leakage current going through the transistors.

Figure 5 shows how shared sleep transistor perform in comparison to non-shared sleep transistor. The width of the shared sleep transistor is equal to the total of the widths of the non-shared sleep transistor thus occupying relatively same area.

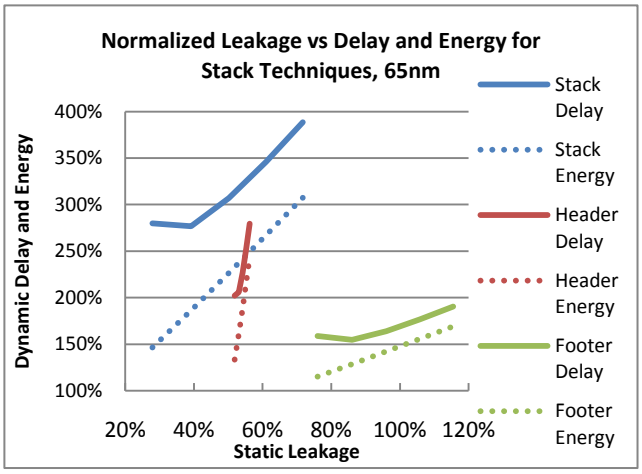

Figure 6: Stack technique performance and impact

We can see how the shared sleep transistor has much less impact on dynamic delay and energy since the transistor is pretty wide, so the current going through it in active mode is significantly bigger thus reducing the dynamic impact of the transistor.

In Figure 6 we show the comparison between three main variations of the stack method. We simulated the method with having stack transistors on both, headers and footers, and only on headers or footers. As in sleep technique, we plotted the information for five different widths of stack transistors. The headers are always twice the size of the footers since PMOS has to be around double the width of NMOS to achieve similar delay characteristics.

We can see how the stack has tremendous amount of dynamic delay and energy overhead since each input has to drive double the number of gates. On the other hand, the leakage is cut down drastically. The leakage reduction this technique reach occurs 100% of the time since the stack technique does not require the circuit to go to sleep mode. This will prove more beneficial in future technologies when the leakage energy is very significant relative to dynamic energy, at that point, we can sacrifice good amount of dynamic energy to achieve high leakage reductions.

Figure 6 shows, as well, how the stack technique reacts changing the width. The dynamic energy impact increases with the increase of width, same as the leakage current. The dynamic delay, however, decreases very slightly at double the minimum width, but it increases again after that. This is opposite to what happen in the sleep technique where the increase of transistor width reduces the dynamic delay impact. This occurs in the stack technique because the increase of the width means extra capacitance to be driven by the input signals. The extra delay that is resulted from that outweighs the delay decrease we get from having more current going through those transistors.

Since stack technique saves significant amount of leakage but have big impact on dynamic delay and energy, the header or footer only stack might prove useful. We can see the leakage is slightly more when using only headers or footers, but the impact is significantly less as well. This might provide a middle ground for the stack technique.

An interesting point we can see in the plot that header only stack is affected by the change of width much less than the footers only stack. The reasoning is because most of the leakage is through the PMOS in the headers since the PMOS has twice the width of the NMOS. This means if we to reduce the leakage through the PMOS we will save most of the leakage. This explains as well why headers stack only has less leakage compared to its counterpart footer.

In order to use the curve by a circuit designer, a horizontal line on the y-axis representing the dynamic delay or energy that the circuit can withstand should be drawn. Then by looking underneath that line, a designer can see what techniques will provide some leakage reduction while staying within the limit of the circuit.

5. CONCLUSION

We have provided the analysis of existing leakage reduction technique using current and predictive technology models. This enabled us to provide some insight whether we need to search for new leakage reduction techniques in the future or keep using the current methods. The analysis also provides grounds to help designers to choose some configuration to meet their requirements.

We show as well how as a result of the increase of the ratio between leakage power and the dynamic power, more drastic techniques that impact the dynamic power drastically will become viable. Stacking transistors on headers and footers is an example of such technique.

For future work we will look at other techniques as well as configurations of the same techniques. Different predictive technology model might be used as well to confirm or invalidate the predictive technology model we used for this paper.

6. REFERENCES

[1] J.C. Park and V.J. Mooney III, "Sleepy Stack Leakage Reduction," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 14, 2006, pp. 1250-1263.

[2] Kyeong-Sik Min, Hun-Dae Choi, H. Choi, H. Kawaguchi, and T. Sakurai, "*Leakage-suppressed clock-gating circuit with Zigzag Super Cut-off CMOS (ZSCCMOS) for leakage-dominant sub-70-nm and sub-1-V-VDD LSIs*," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 14, 2006, pp. 430-435.

[3] Kyung Ki Kim, Yong-Bin Kim, Minsu Choi, and N. Park, "*Leakage Minimization Technique for Nanoscale CMOS VLSI*," *Design & Test of Computers, IEEE*, vol. 24, 2007, pp. 322-330.

[4] N. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. Hu, M. Irwin, M. Kandemir, and V. Narayanan, "*Leakage current: Moore's law meets static power*," *Computer*, vol. 36, 2003, pp. 68-75.

[5] S. Narendra et al., "*Scaling of stack effect and its application for leakage reduction*," in *Low Power Electronics and Design, International Symposium on, 2001.*, 2001, 195-200.

[6] S. Mutoh et al., "*1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS*," *Solid-State Circuits, IEEE Journal of* 30, no. 8 (1995): 847-854.